big data and cognitive computing

MDPI

Article

Spatial Sound in a 3D Virtual Environment: All Bark and No Bite?

Radha Nila Meghanathan [1,*,†], Patrick Ruediger-Flore [2,†], Felix Hekele [1], Jan Spilski [1], Achim Ebert [2] and Thomas Lachmann [1,3]

[1] Centre for Cognitive Science, University of Kaiserslautern, 67663 Kaiserslautern, Germany; felix.hekele@sowi.uni-kl.de (F.H.); jan.spilski@sowi.uni-kl.de (J.S.); lachmann@sowi.uni-kl.de (T.L.)
[2] Human-Computer Interaction Lab, University of Kaiserslautern, 67663 Kaiserslautern, Germany; patrick.ruediger@mv.uni-kl.de (P.R.-F.); ebert@cs.uni-kl.de (A.E.)
[3] Centro de Investigación Nebrija en Cognición (CINC), Universidad Nebrija, 28015 Madrid, Spain
* Correspondence: radha.meghanathan@sowi.uni-kl.de
† Both authors contributed equally.

Abstract: Although the focus of Virtual Reality (VR) lies predominantly on the visual world, acoustic components enhance the functionality of a 3D environment. To study the interaction between visual and auditory modalities in a 3D environment, we investigated the effect of auditory cues on visual searches in 3D virtual environments with both visual and auditory noise. In an experiment, we asked participants to detect visual targets in a 360° video in conditions with and without environmental noise. Auditory cues indicating the target location were either absent or one of simple stereo or binaural audio, both of which assisted sound localization. To investigate the efficacy of these cues in distracting environments, we measured participant performance using a VR headset with an eye tracker. We found that the binaural cue outperformed both stereo and no auditory cues in terms of target detection irrespective of the environmental noise. We used two eye movement measures and two physiological measures to evaluate task dynamics and mental effort. We found that the absence of a cue increased target search duration and target search path, measured as time to fixation and gaze trajectory lengths, respectively. Our physiological measures of blink rate and pupil size showed no difference between the different stadium and cue conditions. Overall, our study provides evidence for the utility of binaural audio in a realistic, noisy and virtual environment for performing a target detection task, which is a crucial part of everyday behaviour—finding someone in a crowd.

Keywords: virtual reality; eye tracking; binaural audio; gaze trajectory; blink rate; sound localization; target detection; visual search

Citation: Meghanathan, R.N.; Ruediger-Flore, P.; Hekele, F.; Spilski, J.; Ebert, A.; Lachmann, T. Spatial Sound in a 3D Virtual Environment: All Bark and No Bite? *Big Data Cogn. Comput.* **2021**, *5*, 79. https://doi.org/10.3390/bdcc5040079

Academic Editor: Min Chen

Received: 31 October 2021
Accepted: 8 December 2021
Published: 13 December 2021

1. Introduction

The feeling of being "immersed" in virtual environments (VEs) has long been an essential element in the design of user experiences for developers and content creators [1]. Virtual environments are immersive when they afford perception of the environment through sensorimotor relationships that mimic our natural existence [2]. The degree of immersion depends on various factors of the visual experience such as the field of view, display latency, display resolution and also the number of other sensory modalities available in the virtual environment. For example, a slowly updated display is less immersive than one that can catch up to the speed of our head movements. In its simplest form, an immersive VE has both visual and auditory modalities [3,4].

Immersive sound in a VE can be achieved by incorporating environmental sounds, sounds of our own actions and simulation of the acoustics of the environment, which will affect the perceived sound [4]. Sound can be incorporated into a VE as simple stereo sounds or as spatial sounds, which render real-world cues such as sound reflections and acoustic changes due to body movements, resulting in the virtual experience being perceived as more authentic. Spatial sounds not only increase the feeling of presence or

'being there' but also elicit more head and body movements from the user on account of being more immersive [5]. However, spatial sounds are complex and both acquisition and reproduction are demanding in terms of the equipment, effort and expense involved. Binaural sound is sound that is perceived as being present in a specific location in space—distance, elevation and azimuth. As the name suggests, it is achieved by simulating how the sound reaches each of our ears. Finding where a sound originates—sound localization, is essential to veridical perception of an environment. High fidelity in spatial sound rendering is uncompromisable since conflicting visual information could interfere with sound localization [6], as commonly seen in the capture effect or ventriloquism effect [7,8]. Serafin and colleagues [4] describe 'ear adequate' headphones, individually administered binaural signals, head movement tracking and room acoustics as some of the requirements of a spatial soundscape to ensure high fidelity in the audio-visual environment. The position of the sound source, the position of the receiver (user), the individual ear and head properties of the receiver, the positions of other objects in the environment and the acoustic properties of the room can all together be used to generate an individualized soundscape. The level of complexity in the type of soundscape incorporated into the VE is application dependent [9]. Therefore, it is more pragmatic and economical to use spatial sounds only when they are effective and add value to the specific VE. For example, spatial sound may not be essential for a virtual lesson with an instructor speaking, whereas, it will be an advantage in a table tennis training environment, where auditory feedback will improve gameplay.

In an investigation of the efficacy of different sound types in a 3D VE, Høeg and colleagues [10] used a visual search task, where the participant was asked to search for a specific visual target randomly positioned in a scene. To assist the user in finding the target, a sound was played to indicate the target location. This auditory cue is akin to a friend calling our name from a crowd, which would help us find them more easily. In this study, the effect of different auditory cues on participant reaction times (RT), that is time to search for the target in the scene, was measured. The authors found that binaurally presented cues facilitated RTs more than stereo cues or the absence of cues by being spatially and temporally synchronous with the visual elements of the display. This finding is significant because it essentially shows that sound localization was better with binaural audio in a virtual environment. However, since the visual environment in this study was a simple 3D visual search display based on a 360° video, it is not clear whether the advantage of a binaural cue will also be present in a more dynamic virtual environment with environmental noise, which is more likely to occur in a real-world setting.

It has been observed that the introduction of noise can obscure audio cues and hinder the detection of visual stimuli [11]. In recent research by Malpica, Serrano, Gutierrez and Masia [12], the introduction of different types of noise led to a severe drop in visual detection and recognition performance in virtual reality (VR), irrespective of the type of noise introduced.

In an inverse effect, sounds went undetected under high perceptual load in the visual modality in an effect known as 'inattentional deafness' [13]. Moreover, visual distractors, even when irrelevant to the task, capture attention [14–17].

The above findings on the effect of auditory and visual noise on perception indicate that behaviour is affected by the presence of noise—both visual and auditory. Therefore, in our study, we tested the efficacy of different types of spatial sounds, specifically, stereo and binaural, in a virtual environment with and without environmental noise. We used a visual search task with different auditory cues to test their relative effects on search performance. This task allowed us to use both visual and auditory modalities in the VE and to place our task at the intersection of the two modalities. In this manner, we studied both visual target identification and sound localization simultaneously in a noisy virtual environment. A sound localization task or a visual search task on their own would be insufficient to understand perception of auditory and visual stimuli in an ecologically valid virtual environment. Our setup enabled the study of the interaction between both

visual and auditory modalities in a VE. In this manner, we mimicked a common scenario in everyday life—looking for someone in a crowd, which is made easier if they call to us. This is where the advantage of binaural audio—that it can be placed at a distance and elevation along an azimuth—comes into play. The sound source would be congruent with the visual stimulus location enhancing stimulus detection. A stereo cue, in contrast, only has slight delays between the inputs to the left and the right ear, which gives the illusion of depth, but does not enable accurate sound localization. Therefore, the binaural audio cue is expected to facilitate visual search more than the stereo cue, as already found in previous literature [10,18]. This will not necessarily be true in the condition with environmental noise, where both auditory and visual distractors will interfere with target search and localization.

The interim results from our study have already been described elsewhere [19]. The descriptive results indicated lower performance variability in the presence of an auditory cue with a slight indication that the binaural cue may be more advantageous than the stereo cue. Participants did not report differences in mental load between the experimental conditions on any dimension measured using the NASA-TLX questionnaire [20]. Participants generally reported high spatial presence in the task as measured using the Igroup Presence Questionnaire (IPQ) [21].

In the present investigation, we derived four measures from the eye tracking data we collected during the experiment—two measures pertaining to the spatiotemporal characteristics of eye movements made during the task and two measures pertaining to the physiological response to the task. We chose these measures to obtain a comprehensive understanding of behaviour in a rich virtual environment. These measures would tease apart the different cognitive processes that contribute to task performance, allowing us to assess the effect of the environment and the different auditory cues.

The first measure, time to first fixation (*TFF*), quantifies the time for target search, which is an indicator of the speed of target localization. The second measure, gaze trajectory length (*GTL*), quantifies the length of the search path, which gives us insight into the search process adopted by participants. The *TFF* results are expected to replicate the results obtained by Høeg et al. [10]. We expect that the binaural cues will result in shorter search times (*TFF*) and shorter search paths (*GTL*) than the stereo and no cue cases in the noise-free environment. Such a result would indicate quicker target detection with binaural cues in a realistic environment, which would strengthen the case for binaural sound use in VEs.

We do not have a specific prediction about whether the same results will be obtained in the condition with environmental noise. Even if the cues are effective, the presence of distracting noise could make the search task more effortful. In our interim analysis, although there was no discernible pattern in the mental effort report of participants, there was a report of frustration in the conditions with environmental noise [19]. Therefore, in the present study, we focused on two measures of mental effort that could be derived from the eye tracking data-blink rate and pupil size. Pupil diameter is a well-established indicator of cognitive load that increases with increase in load [22–25]. It has been tested as an indicator of mental effort in practical applications such as combat [26], driving [27,28] and surgery [29]. In contrast, blink rate is a more ambiguous measure. In some studies, blink rate has been reported to decrease with cognitive load [30], while in others, blink rate has been reported to increase with load [24]. A more complicated relationship of blink rate with different types of loads has been found in other studies [31,32]. In our study, we expect pupil size to increase in the conditions with environmental noise, while no specific prediction is made for blink size. We also expect pupil size to be higher in conditions without the cue than with stereo or binaural cue. An advantage for binaural cue in terms of cognitive load measures would be the ultimate benchmark for the utility of binaural sounds in VEs.

2. Materials and Methods

The dataset used in this article was obtained from a VR experiment. The stimuli used, pre-tests for stimulus validation, study setup and description of the data obtained in the experiment are detailed in Ruediger et al. [19].

2.1. Participants

A total of 20 participants (8 female) from the University of Kaiserslautern aged between 22 and 32 years (M = 27.32; SD = 2.97) volunteered to perform the experiment with informed consent. Most participants reported relatively little previous experience with virtual reality on a 5-point scale (M = 2.73; SD = 0.86) ranging between 'First time use' and 'Already living in VR'. Data from two participants, whose data were not recorded in one or more experimental conditions due to technical errors, were removed from the analysis. One more participant with extreme values was removed from the analysis as explained in Section 2.5.2. Data from the remaining 17 participants were analysed.

2.2. Stimuli

The VR stimuli were presented in an HTC Vive with an integrated Tobii eye-tracker. Each stimulus consisted of a scene acquired using simultaneous 360° video and audio recording of a real-world handball game stadium. Both the scene and the scene acquisition method were selected in order to achieve maximum fit with reality. A sport environment integrated rich visual and auditory information in the scene. Moreover, the simultaneous video and audio recording ensured that auditory noise was spatially synchronized with the visual stimulus. This synchrony ensured that the acoustic cues were separable from the environmental noise.

There were two stadium conditions: empty and full. In the empty stadium condition, the scene included a video of the empty stadium with sparse activity from groundskeepers, etc., resulting in low visual and auditory background noise. In the full stadium condition, the scene included a live audience in the stadium and players entering the handball court, which resulted in a condition with a noisier visual and auditory background.

Besides the two stadium conditions, there were three auditory cue conditions: no cue, binaural cue, stereo cue. The auditory cue was an air horn signal, which is a typical sound at handball games, as horn signals are used by fans to cheer for the team. In addition, this horn signal was evaluated in a pre-test to ensure that participants were able to distinguish the stereo and binaural cues [19]. The sound for the cue was rendered using the Adobe Premiere Toolbox in both stereo and binaural using the generic standard head related transfer function. The combination of the two stadium and three cue conditions resulted in six scenes (videos).

The visual target presented in the scene was a set of three Minions (fictional, yellow creatures from the popular movie franchise) stacked vertically (Figure 1). The minion was chosen as a visual search target to mimic the problem of finding a person in a crowd in a VR setting. These targets blend in with the yellow home team shirts, but at the same time, are easily recognizable because of their unique and broadly known appearance. The colour scheme of the visual stimuli ensured they could not be identified in the peripheral visual field; however, they were distinct enough to be quickly identifiable once focused. These targets were presented in one of six locations on the azimuth plane ($-135°$, $-90°$, $-45°$ and $45°$, $90°$, $135°$) for each stadium and cue condition, resulting in 36 trials (2 stadia $\times$ 3 cues $\times$ 6 locations). A pre-test was conducted to determine the number of recognizable directions of a sound cue for the chosen signal. The pre-test revealed that only changes in the azimuth plane were perceived correctly. Therefore, no variations in the elevation were made in the experimental setting.

Figure 1. Fixation behaviour from one participant in the empty binaural (**top**) and another participant in the full binaural (**bottom**) conditions overlaid on a snapshot of a trial. Fixations are indicated by purple and red circles. The area where the central fixation cross was presented is indicated by a black square. The six locations are indicated as Tx (T1 to T6), where x is the location's position in the sequence of six target presentations in a condition. Participant fixations are concentrated over the fixation cross and also in the six target locations. Background is partially scrambled to remove advertisements and faces are blurred to hide identities.

2.3. Data Acquisition

Data was acquired at 100 Hz sampling frequency using a Tobii Pro VR Integration eye-tracker retrofitted to an HTC Vive (2016 version). Data from each eye was recorded via one eye-tracking sensor and ten infrared illuminators with a total trackable field of view of 110° for eye movements.

2.4. Procedure

Participants were instructed to perform a search task in a VR environment. After the participant was seated in the lab, they wore the VR headset and their eyes were calibrated using a 9-point calibration. The instructions for performing the task were displayed on the screen before the participant was presented with the six experimental trials for one stadium-cue combination. Each trial began with a large, blue cross in the centre of a stadium scene, presented for approximately five seconds. The participant was asked to fixate on the blue cross and begin searching for the target as soon as the blue cross disappeared. In the two conditions with auditory cues, cue presentation was synchronized with the disappearance of the blue cross. Participants had been instructed to look around the stadium scene and find the visual target, which was presented for 6840 ms from trial onset. On finding the visual target, they were asked to fixate on the target until it disappeared. The end of the trial was indicated by two, large red arrows directing the participant toward the central blue cross. After these 6 trials, the participant was asked to answer the NASA-TLX questionnaire [20]. The NASA-TLX is a well-established tool to provide a reliable assessment of perceived mental workload [20,33]. A 20-level version of the NASA-TLX was used for this study (1 = low to 20 = high).

The above procedure was done for all six stadium × cue combinations. All six target locations were presented in random sequence for each of the six stadium-cue combinations, which were also presented in random order for each participant. After the end of all trials, the participant was also asked to fill in the IPQ [21]. The IPQ consists of 14 items with a seven-point Likert scale (0 to 6). The 14 items load on the four factors Spatial Presence (SP), General Presence (GP), Involvement (INV) and Experienced Realism (REAL). We observed [19] that SP was rather high, while the single item factor GP, INV and REAL were expectedly lower. The entire experiment, including the 36 trials, the six NASA-TLX questionnaires and the IPQ, took approximately 30 min to complete.

2.5. Data Analysis

Data were analysed only from those trials where the visual targets were successfully found. Target hits were assessed using predetermined areas of interest, which were superimposed on the stimulus material via Tobii Pro Lab v1.152 software (Tobii AB, Danderyd, Sweden). If the visual target was fixated on for at least 70 ms, the fixation was considered a target hit and the trial was included in the analysis. Trials with extreme values, as explained below, were also removed from the analysis, such that the same set of trials were analysed for each measure. The fixation measures obtained using Tobii Pro Studio and all the other measures were computed using custom-written Python scripts. Only the results of eye-tracking data are analysed and described below. The results from the NASA-TLX and IPQ have been described in Ruediger et al. [19].

2.5.1. Time to First Fixation (*TFF*)

The *TFF* was calculated as the time between the appearance of the visual target and the first fixation on the target using Tobii Pro Studio. The *TFF* was considered an indicator of search performance. However, during the experiment, experimenters observed participant differences in the speed at which the task was performed. Since participants were not given any instructions regarding speed for performing the task, participants may have adopted a slower or faster pace at which to perform the task. This can be observed in the high *TFF* variances (see boxplot Figure 2a), which might render it incomparable between participants. Therefore, we computed an additional measure of the search path as described in Section 2.5.2.

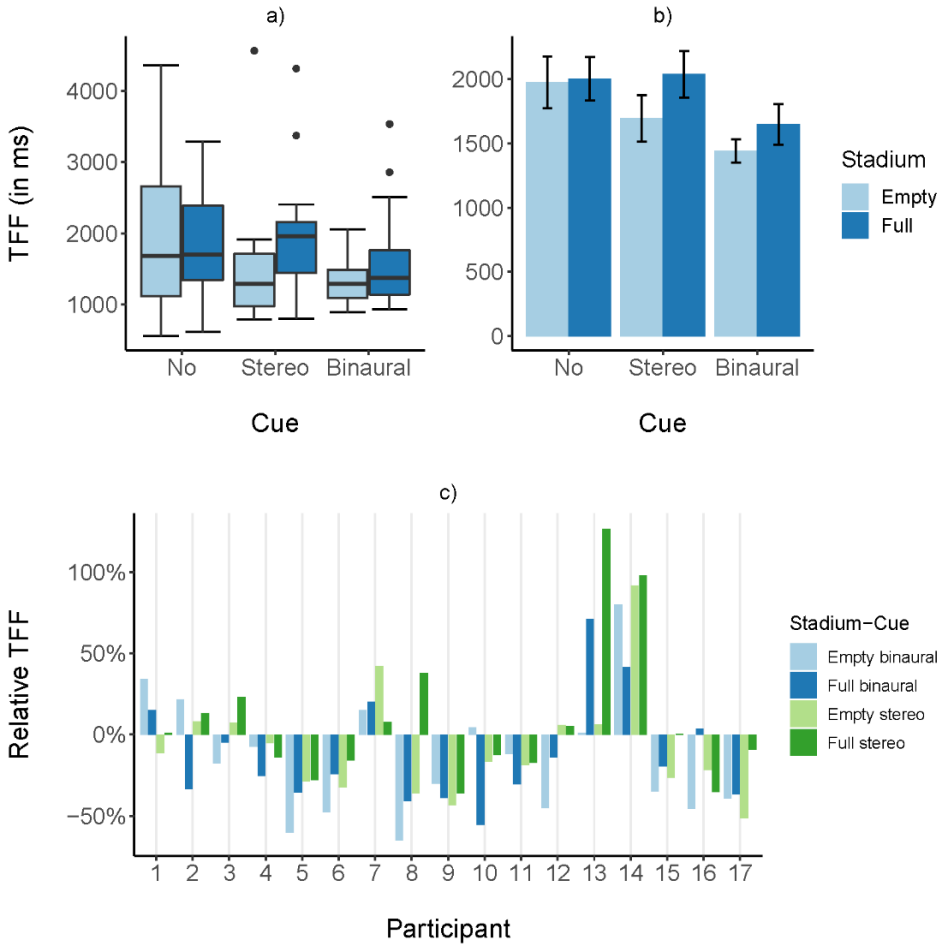

Figure 2. (**a**) Boxplot with median *TFFs* and (**b**) barplot with mean *TFFs* and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. (**c**) Relative changes in *TFF* for 17 participants in no cue, stereo and binaural cue conditions for empty and full stadium conditions. The no cue condition is interpreted as a baseline for the stereo and binaural conditions.

For the analysis, trials with *TFF* with extreme values (less than 50 ms) were removed since the search interval was not reliable. This resulted in 4% data loss (3 trials from 2 different participants lost).

2.5.2. Gaze Trajectory Length (*GTL*)

Gaze trajectory length was calculated by adding up the pairwise differences between normalized gaze point coordinates $G\,(x, y)$ of subsequent timestamps recorded until the first fixation on the target occurred (after n time steps).

$$GTL = \sum_{i=1}^{n} (G_{i-1} - G_i)$$
$$GTL = |GTL|$$

Since the gaze points were normalized with respect to the extent of the scene, *GTL* values are a factor of the scene width. For example, a *GTL* value of 2 implies that the gaze path was twice the scene width.

GTL is an indicator of the search path adopted by the participant. With easier search, the *GTL* would be shorter, whereas, in more prolonged search trials where the participant searches in more locations on the scene, *GTL* would be longer. The *GTL* suffers from the same individual differences as the *TFF*. As a consequence, reliable comparisons are only

possible within each participant or by introducing a relative dimensionless measurement, e.g., the ratio of the measures between the different conditions relative to the no cue condition as depicted in Figures 2c and 3c.

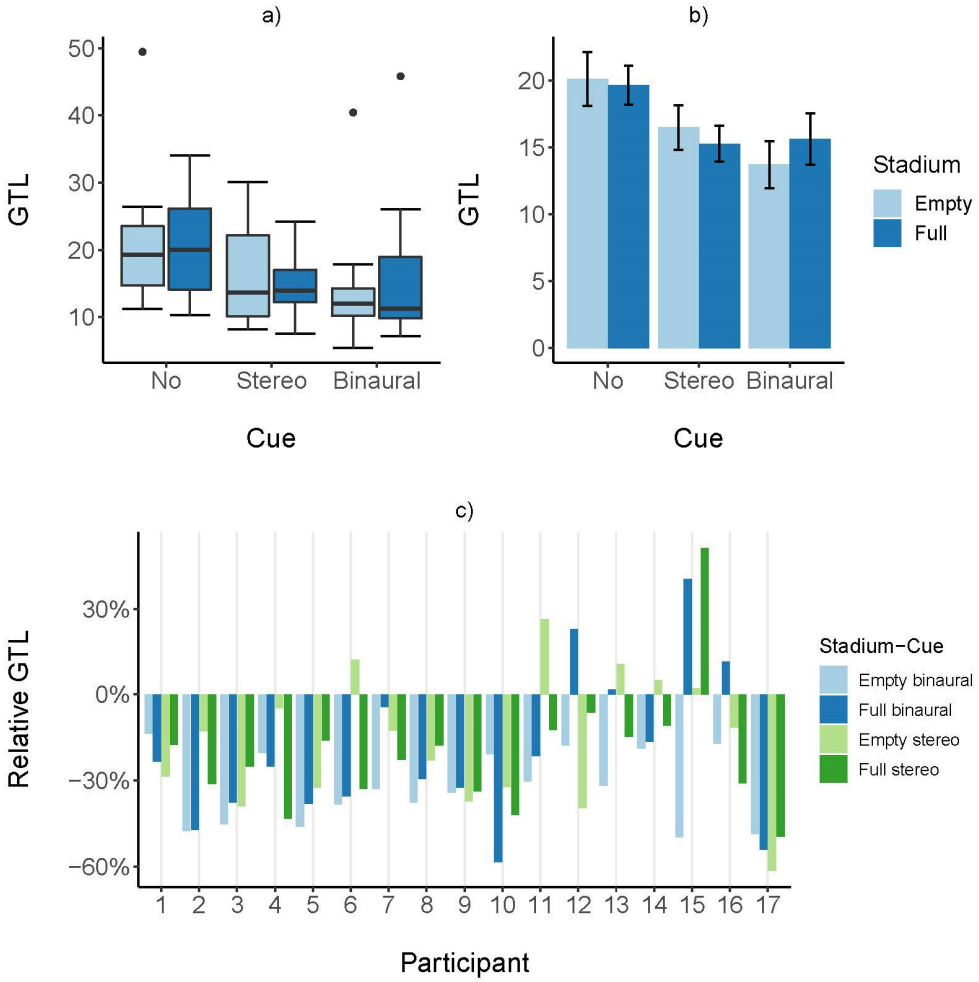

Figure 3. (**a**) Boxplot with median *GTLs* and (**b**) barplot with mean *GTLs* and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. *GTL* is measured in terms of the number of scene widths spanned. (**c**) Relative changes in *GTL* for 17 participants in no cue, stereo and binaural cue conditions for empty and full stadium conditions. The no cue condition is interpreted as a baseline for the stereo and binaural conditions.

Data from one participant with extreme gaze trajectory lengths was removed from the analysis. The extremely long search paths appeared to be due to a technical error.

2.5.3. Blink Rate

The blink rate during search was used as a measure of cognitive effort. To identify blinks, the pupil size data stream from the eye recording was used. Blinks are represented as missing values in the pupil data. However, the pupil value could also go missing because of other small eye movements, measurement artefacts, etc. Therefore, the pupil size data was first preprocessed by identifying small artefacts as missing values of 50 ms or less. These values were filled with the last valid pupil value. On this artefact-corrected pupil series, blinks were identified using the algorithm devised by Hershman et al. [34]. The algorithm identifies the correct start and end of the blink by identifying a decrease preceding and an increase succeeding a sequence of missing pupil values. Blinks that occur within 50 ms of each other are also merged into one larger blink. However, the pupil value

could also go missing because of other small eye movements, measurement artefacts, etc. Therefore, for the purpose of obtaining the blink rate, missing values were identified as blinks only when the blink duration was greater than 100 ms. After this step, the number of blinks was counted for each trial from the trial onset until the first fixation on target. Finally, blink rate was calculated as

$$Blink\ rate = \frac{Blink\ count}{TFF}$$

2.5.4. Pupil Size

Like blink rate, pupil size was also used as a measure of cognitive load. For this purpose, the pupil size data series marked with the blinks identified in the previous step was used. After identification of blinks, irrespective of blink duration, blink regions were interpolated using an order-3 spline 100 ms before and after the blink. Using this interpolated series, we calculated baseline-corrected average pupil size from trial onset until the first fixation on the target. Although there was a potential baseline interval of no activity when the blue cross was presented, it could not be used because participants did not always fixate the cross steadily. Therefore, the mean pupil size for each participant across all conditions was calculated as the baseline pupil size. This baseline was subtracted from the mean pupil size in each condition and target location giving us the demeaned Pupil Size.

2.5.5. Statistical Analysis

The data was analysed using repeated-measures ANOVA with stadium and cue as factors. For pupil size alone, an additional analysis was performed for the empty stadium trials with cue as a single factor. In case of violation of sphericity, the Greenhouse–Geisser corrected p-values are reported. For post hoc tests with multiple pairwise comparisons, Tukey-adjusted p-values are reported.

3. Results

3.1. Hit Rate

We assessed search performance by performing a repeated-measures ANOVA on target hit rate with stadium and cue as factors. There was a significant effect of cue on target hit rate, F (2, 32) = 3.3, p = 0.049, η^2_G = 0.032. Post hoc tests showed that the binaural cue conditions had a higher hit rate than the no cue conditions (p = 0.04) averaged over the empty and full stadium conditions as shown in Table 1.

Table 1. Mean target hit rate (and standard deviation) for the two stadium conditions (empty and full) and three cue conditions (binaural, stereo, no cue) for 17 participants.

	No Cue	Stereo Cue	Binaural Cue
Empty stadium	92 (15.7)	97 (6.5)	96 (12.5)
Full stadium	91 (17.8)	94 (14.4)	98 (5.5)

3.2. Time to First Fixation (TFF)

As mentioned earlier, there was large variability in the *TFF* across participants, especially in the no cue condition (Figure 2a). An ANOVA performed on the mean *TFF* values revealed a significant effect of cue, F (2, 32) = 7.5, p = 0.003, η^2_G = 0.07. Post hoc tests showed lower *TFF* in the binaural cue condition (Figure 2b) than in the stereo (p = 0.03) and no cue conditions (p = 0.002) averaged over the stadium conditions.

While the measure *TFF* is generally highly objective, it suffers from individual participant effects, such as training or experience in related tasks or orientation in virtual reality in general. As a consequence, it might be erroneous to only evaluate the mean values for 17 participants. Therefore, we additionally investigated the *TFFs* for each participant to

understand individual differences. For each participant, we calculated the relative change ratios between each of the cued conditions with respect to the no cue condition for each stadium condition and participant as depicted in Figure 2c. We found that search times were lower in the empty than the full stadium conditions for 12 participants in the binaural cue conditions and for 11 participants in the stereo cue conditions. For 14 participants, the binaural cue showed lower *TFFs* than the stereo cue conditions, while one participant showed the opposite effect. For the remaining two participants, the difference did not show a clear pattern.

3.3. Gaze Trajectory Length (GTL)

There was a significant effect of cue on *GTL*, F (2, 32) = 16.6, $p < 0.001$, $\eta^2_G = 0.09$. Post hoc tests for values averaged over both stadium conditions showed a higher *GTL* (Figure 3) in the no cue condition than in the stereo and binaural cue conditions (both $p < 0.001$). Here, again, we computed *GTL* as a percentage change relative to the no cue condition. We observed that for most participants, *GTL* decreased compared to the no cue condition (Figure 3c), which confirmed our findings from the ANOVA.

3.4. Blink Rate

We observed large within-subject and between-subject variability in blink rates. Some trials had no blinks at all, while others had a very large number of blinks (Figure 4a). The binaural cue condition had the least number of blinks. Across cue conditions, three were fewer blinks in the empty stadium condition than the full stadium condition.

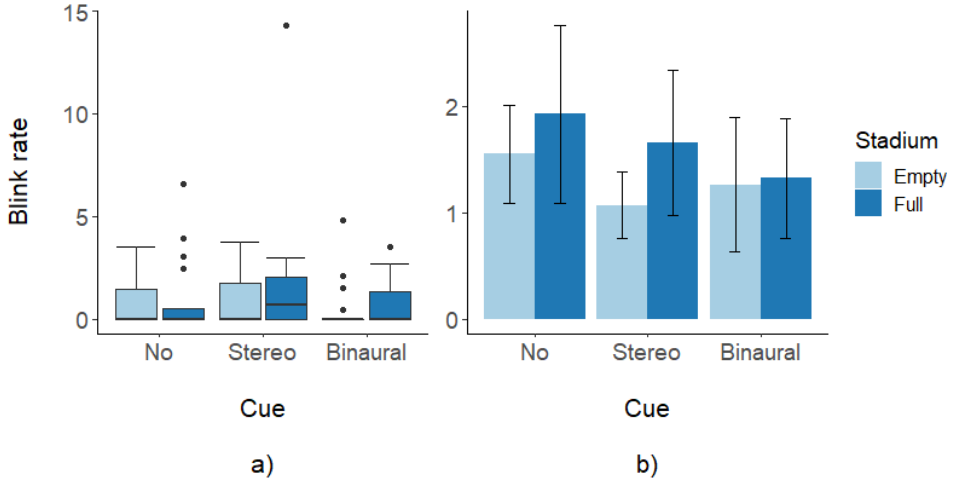

Figure 4. (**a**) Boxplot with median blink rates and (**b**) a barplot with mean blink rates and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions. Blink rate is the number of blinks made per second.

A 2 × 3 repeated measures ANOVA on mean blink rates did not show a significant effect of stadium or cue (Figure 4b).

3.5. Pupil Size

ANOVA on average demeaned pupil size revealed a significant effect of stadium, F (1, 16) = 53.3, $p < 0.001$, $\eta^2_G = 0.597$, with higher pupil size in the full stadium than in the empty stadium condition ($p < 0.001$). There was also a significant interaction between stadium and cue, F (2, 32) = 5.5, $p = 0.01$, $\eta^2_G = 0.067$). Post hoc tests revealed a higher pupil size in the full stadium condition for all cue conditions ($p < 0.001$), although there was no effect of cue itself in any of the stadium conditions (Figure 5).

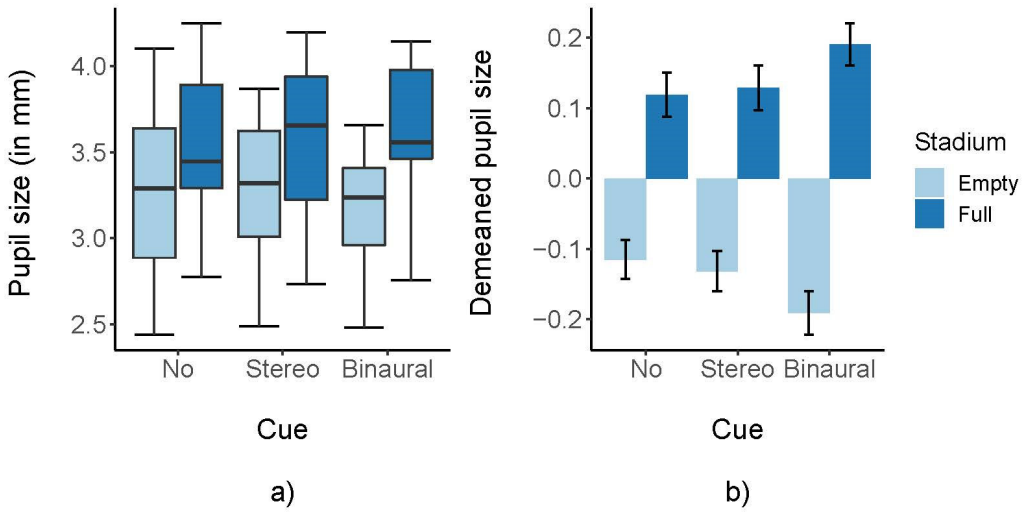

Figure 5. (**a**) Boxplot with median pupil sizes and (**b**) a barplot with demeaned pupil size and error bars indicating standard errors of mean for 17 participants in the no cue, stereo and binaural cue conditions for the empty and full stadium conditions.

However, pupil size is sensitive to luminance variations. Since the empty and full stadium conditions are visually different, the full stadium condition has much larger luminance variations, which might affect pupil size differently, as seen in the pupil size deviations from mean. Therefore, we assessed pupil size again only in the empty stadium condition with only cue as a factor. This analysis did not show a significant effect of cue on pupil size.

4. Discussion

In this study, we compared the effect of three types of auditory cues (no cue, stereo cue and binaural cue) on visual search behaviour in two types of virtual environments using four measures—time to first fixation (*TFF*), gaze trajectory length (*GTL*), blink rate and pupil size.

4.1. Behavioural and Eye Position Measures

First, we found a performance advantage for binaural cue in comparison to trials where cue was absent, whereas, such an advantage was not present for stereo cues. This improved performance was also visible in the target search duration (*TFF*). Participants were quicker to find the target with the help of a binaural cue than with a stereo cue or in the absence of an auditory cue (Section 3.2). This result is in line with the quicker search times obtained in the studies by Hoeg and colleagues [10] and Brungart and colleagues [18].

The gaze trajectory length (*GTL*) measure, which quantified the length of the search path (Section 3.3), revealed a cue advantage as well. Trials with no auditory cue showed longer search paths than trials with binaural and stereo cues, clearly showing a benefit of the auditory cue. However, there was no difference between the search paths of the stereo and binaural cues.

Although not statistically significant, the boxplots and summary barplots (Figure 2a,b) show that, in the presence of a cue, search durations (*TFF*) were higher when the stadium had distractors (full stadium condition) than when it did not (empty stadium condition). This effect may be attributed to distracted search in the full stadium only in the presence of auditory cues, since such an effect is not present when there is no auditory cue. This is visible in the individual participant data (Figure 2c), where 12 of 17 participants show lower search times in empty than full stadium conditions for the binaural cues (11 for stereo cues). While research combining task performance with fully moving VEs is scarce, related research could provide additional insight. Olk and colleagues [35] reported slower detection of stimuli in a VE when those stimuli were harder to distinguish from surrounding

objects either due to their distance or distinctiveness. This indicates that the minions, which were chosen to merge with the yellow elements of the VE, were indeed less distinctive when the players and audience with yellow uniforms appeared in the full stadium condition. Moreover, the appearance of a person—a strong social element—was recently shown to influence participants' visual attention in virtual reality [36], where a person was fixated significantly more in a 360° video compared to a 2D video. In our task, the players and audience in the full stadium condition would have similarly attracted attention. This validates the minions as an appropriate visual target for our task. Additionally, the presence of people, even though task-irrelevant, also negatively impacted task performance in our study.

However, no such difference is seen between the two stadium conditions for the search paths. To understand this disparity, we additionally investigated *TFF* and *GTL* by separating the trials based on the location of the targets (Figure 6). For *TFF*, in the empty stadium, we found a high association between the target distance from the centre (in degree) and the time to fixation of the target. For the full stadium, this association holds true only in the presence of a binaural cue and to a lesser degree in the presence of a stereo cue. This implies that the full stadium interferes with the search process as expected.

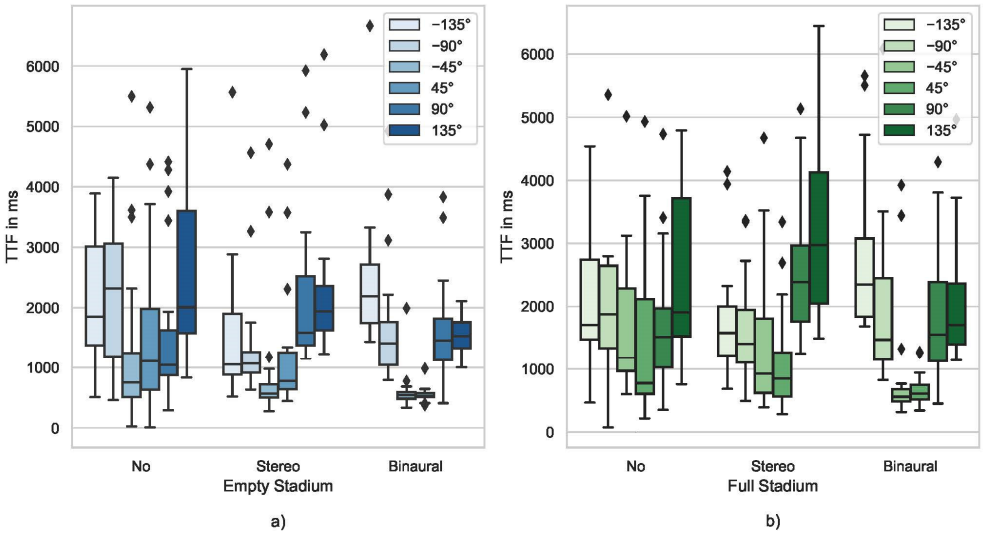

Figure 6. Boxplots of *TFF* for no cue, stereo and binaural cue conditions for empty (**a**) and full stadium (**b**) conditions grouped by target location (search trials). The target locations are marked by their relative angle to the starting position.

In contrast, for the gaze length trajectories, we could not find such a pattern. It remains unclear how the eye and head movements necessary for different target positions in the different stadium conditions moderate the overall results. This could have been because of a design shortcoming. As mentioned earlier, our participants did not always fixate exactly on the large, central blue cross before the start of each trial. This distracted trial beginning could mean that the search would not have always started from the middle of the display, leading to inconsistent *GTLs*. One solution to this problem would be to force the participant to fixate on the central cross to begin the trial. Alternatively, a more dynamic setup would have averted this problem by presenting the cue depending on the participant's current gaze location.

Another source of inconsistency in the results was individual differences. Comparing the per participant effects in Figures 3 and 4, the individual measure varies strongly for both gaze paths (*GTL*) and search duration (*TFF*). We could not spot any general pattern describing the relative degree of changes in either of the measurements. Larger sample sizes are required in future studies to mitigate the effect of this variability. Any variability stemming from differing levels of familiarity with virtual reality technology, although low

as mentioned in Section 2.1, can also be explored with a larger sample size. In spite of the individual effects heavily moderating the degree between the absence and presence of an auditory cue in this visual search task, we found that the presence of any auditory cue speeds up the search performance.

Overall, the search duration (*TFF*) and search path (*GTL*) measures presented to be useful metrics of search behaviour in our task. Together, they have revealed a search advantage of auditory cues, with the binaural cue being slightly more advantageous than the stereo cue.

4.2. Physiological Measures

The next two measures we tested—blink rate and pupil size—were physiological measures, both of which have been studied in virtual environments. Although not statistically significant, we found lower blink rates in the empty stadium conditions than in the full stadium conditions. However, the trials without cues did not show such an effect. On the contrary, we observed lower blink rates in the easier trials with the binaural cue. Blink rate is an inconclusive measure of cognitive effort [37]. Blink rate is known to decrease in cases of extreme focus and increased workload, as observed in surgeons [30,38]. Veltman and Gaillard [30] indicate a distinction in the underlying factors that affect blink rate. They found that blink rate decreased when more visual information had to be processed, while it increased when the difficulty of the task increased. In an experiment systematically varying visual and cognitive demands, Recarte and colleagues [31] found that blink rate decreased with visual load and increased with mental load. In a driving task, Merat et al. [32] found a similar fall in the blink rate with increased visual information in the absence of a secondary task. Adding a secondary task increased blink rates, although some results did not fit this pattern, indicating a tradeoff in blink behaviour between visual information intake and mental workload. These U-shaped patterns in blink rates have been interpreted differently by others. Berguer and colleagues [39] found that surgeons had lower blink rates when performing surgery than at rest, but blink rate increased while doing the same in a laparoscopic environment. They interpreted this result as the outcome of a conflict between task demand or stress and concentration. Zheng et al. [38] found, in a VR laparoscopic surgery setting, that those participants who reported more frustration and mental effort in the NASA-TLX blinked less frequently. It is also worth noting that some studies have not reported an effect of mental load on blink rate, while pupil size or other measures responded to load [28,40].

In the context of the ambiguous nature of factors affecting blink rate that were discussed above, our results did not show a discernible pattern to draw parallels to any of the above literature. Although the full stadium condition had higher visual information in the display, this information was task-irrelevant, and therefore, it cannot be equated to the demand of having to process additional visual information as described above. Our task may have been too easy to elicit an effect of visual or mental load in comparison to the difficult driving and surgery scenarios that have been studied. In addition, the median blink rates and an investigation of individual participant blink rates revealed high variability in the data. Large variability in blink rate was also reported by Benedetto et al. [28]. In spite of blink rate being decreased in head-mounted VR displays in comparison to monitors or natural settings [41], in our data, some participants showed extremely large blink rates (up to 15), which may indicate poor data quality. Blinks are identified when the pupil is not detected by the eye tracker. The Tobii Glasses eye tracker we used was embedded in the VR headset, which should have resulted in lesser data loss. However, loss of the pupil size data stream occurred more frequently for some participants (as high as 19% for one participant). Some participants wore glasses and/or lenses, which may have resulted in higher data loss. This is a shortcoming of video-based eyetracking, which needs to be overcome to increase the reach of eye tracking integrated VR setups. A simple solution to this problem would be to record a video of the participant's eyes, which would allow us to manually identify blinks.

The comparability of our results with existing literature is additionally made difficult by the fact that the blink sensors (remote eye tracker, head-mounted eye tracker, EOG) and blink detection algorithms (manual, automatic, different duration thresholds, etc.) are all different. If future studies report data quality and the precise parameters used for blink detection, it will become easier to reach a consensus on this complex measure.

Our last measure, pupil size, showed only an effect of the stadium with larger pupil sizes in the full stadium condition. This effect is clearly due to luminance differences between the two scenes. Pupil size responds to both changes in luminance and cognitive effort [25] and our result shows that the task-evoked pupillary response (TEPRs) was not separable from luminance effects. However, even in the empty stadium trials, where we can reasonably assume equivalent luminance between cue conditions, any increase in cognitive effort that may have been present was not seen in our results except as a small decrease in median pupil size. It should be noted that TEPRs are small changes that require a large number of trials to be averaged and reflect large changes in cognitive load [22,23], both of which did not apply to our study.

Although our scene was chosen to be visually and auditory realistic, which was an advantage for immersion and presence in VE, the realistic stimulus was also partly the reason why the physiological measures did not perform well. We could conduct the same study with more fine-grained control over the visual and auditory noise levels in the environment. The experiment could have only auditory noise or only visual noise as additional conditions to isolate the effect of noise from the amount of visual input that needs to be processed. This would enable the use of both pupil size and blink rate.

For future use of pupil size in our paradigm, besides correcting the design constraints mentioned above, technical sources of error need to be accounted for as well. Eye movements themselves cause distortions in pupil size, which are most evident in different camera viewing angles. Correcting these distortions requires complex mathematical models [42]. Most eye trackers incorporate these perspective-distortion corrections; however, individual differences might still exist (described and modelled in Mathur et al. [42]). One solution is to use measures that are resistant to luminance changes and pupil distortions, such as the Index of Pupillary Activity [43], which measures changes in the oscillatory behaviour of pupil data. However, this measure requires long trial durations, which was not the case in most of our trials.

4.3. Eye Tracking and Immersive VR

Eyetracking gives us access to many dimensions of behaviour. It extends the study of simple behavioural responses by giving us a more fine-grained insight into human interaction with the environment. In our task, we could have asked participants to simply press a button on detecting a target. However, recording eye movement data instead, allowed us to look more closely into the search strategy of each participant. It also allowed the participant to perform the task more naturally without having to remember buttons that they might usually have to press. Such eye movement paradigms make stimulus-response paradigms more seamless and ecologically valid. However, as discussed above, some of the measures obtained from eye tracking data have shortcomings that need to be overcome. Higher fidelity signals will be required in the future for effective use of systems that can provide metrics of cognitive effort for improving user experience and for providing user feedback.

In spite of the lack of results from the physiological measures, our eye position measures have revealed a definitive advantage of an auditory cue for target localization and detection in a virtual environment. We also found that visual and auditory noise interfered with target localization in the presence of facilitating cues. There is also an indication of the usefulness of the binaural cue, which was seen in spite of the large individual differences between the different degrees of environmental noise. This evidence is in support of the use of spatial sound in different virtual environments to improve responsiveness and immersion in the environment. Our study can be extended in future

research with different environments and different degrees of noise to obtain a more comprehensive understanding of sound localization and perception in realistic VEs. This would enable the design of more effective virtual environments with appropriate use of binaural sounds.

Author Contributions: Conceptualization, P.R.-F. and J.S.; Methodology, R.N.M. and P.R.-F.; software, R.N.M. and P.R.-F.; validation, R.N.M. and P.R.-F.; formal analysis, R.N.M. and P.R.-F.; investigation, R.N.M. and P.R.-F.; data curation, R.N.M., P.R.-F. and F.H.; writing—original draft preparation, R.N.M.; writing—review and editing, P.R.-F., F.H., J.S. and T.L.; visualization, R.N.M., P.R.-F. and F.H.; supervision, J.S., T.L. and A.E.; project administration, T.L. and A.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Grassini, S.; Laumann, K.; Rasmussen Skogstad, M. The Use of Virtual Reality Alone Does Not Promote Training Performance (but Sense of Presence Does). *Front. Psych.* **2020**, *11*, 1743. [CrossRef]
2. Slater, M.; Sanches-Vives, M.V. Enhancing Our Lives with Immersive Virtual Reality. *Front. Robot. AI* **2016**, *3*, 74. [CrossRef]
3. Slater, M.; Sanches-Vives, M.V. Transcending the Self in Immersive Virtual Reality. *Computer* **2014**, *47*, 24–30. [CrossRef]
4. Serafin, S.; Geronazzo, M.; Erkut, C.; Nilsson, N.C.; Nordahl, R. Sonic Interactions in Virtual Reality: State of the Art, Current Challenges, and Future Directions. *IEEE Comput. Graph. Appl.* **2018**, *38*, 31–43. [CrossRef]
5. Nordahl, R.; Nilsson, N.C. The sound of being there: Presence and interactive audio in immersive virtual reality. In *The Oxford Handbook of Interactive Audio*; Collins, K., Kapralos, B., Tessler, H., Eds.; Oxford University Press: Oxford, UK, 2014.
6. Jackson, C.V. Visual Factors in Auditory Localization. *Q. J. Exp. Psychol.* **1953**, *5*, 52–65. [CrossRef]
7. Bertelson, P. Starting from the ventriloquist: The perception of multimodal events. In *Advances in Psychological Science, Biological and Cognitive Aspects*; Sabourin, M., Craik, F., Robert, M., Eds.; Psychology Press/Erlbaum: London, UK, 1998; pp. 419–439.
8. Alias, D.; Burr, D. The Ventriloquist Effect Results from Near-Optimal Bimodal Integration. *Curr. Biol.* **2004**, *14*, 257–262. [CrossRef] [PubMed]
9. Larsson, P.; Väljamäe, A.; Västfjäll, D.; Tajadura-Jiménez, A.; Kleiner, M. Auditory-Induced Presence in Mixed Reality Environments and Related Technology. In *The Engineering of Mixed Reality Systems*; Dubois, E., Gray, P., Nigay, L., Eds.; Human-Computer Interaction Series; Springer: London, UK, 2010; pp. 143–163.
10. Hoeg, E.R.; Gerry, L.J.; Thomsen, L.; Nilsson, N.C.; Serafin, S. Binaural sound reduces reaction time in a virtual reality search task. In Proceedings of the 2017 IEEE 3rd VR Workshop on Sonic Interactions for Virtual Environments (SIVE), Los Angeles, CA, USA, 19 March 2017.
11. Hidaka, S.; Ide, M. Sound can suppress visual perception. *Sci. Rep.* **2015**, *5*, 10483. [CrossRef]
12. Malpica, S.; Serrano, A.; Gutierrez, D.; Masia, B. Auditory stimuli degrade visual performance in virtual reality. *Sci. Rep.* **2020**, *10*, 12363. [CrossRef]
13. Macdonald, J.S.; Lavie, N. Visual perceptual load induces inattentional deafness. *Atten. Percept. Psychophys.* **2011**, *73*, 1780–1789. [CrossRef]
14. Theeuwes, J. Stimulus-driven capture and attentional set: Selective search for color and visual abrupt onsets. *J. Exp. Psychol. Hum. Percept. Perform.* **1994**, *20*, 799–806. [CrossRef]
15. Lavie, N. Attention, Distraction, and Cognitive Control under Load. *Curr. Dir. Psychol. Sci.* **2010**, *19*, 143–148. [CrossRef]
16. Forster, S.; Lavie, N. Entirely irrelevant distractors can capture and captivate attention. *Psychon. Bull. Rev.* **2011**, *18*, 1064–1070. [CrossRef]
17. Lavie, N.; Dalton, P. Load Theory of Attention and Cognitive Control. In *The Oxford Handbook of Attention*; Nobre, A.C., Kastner, S., Eds.; Oxford University Press: Oxford, UK, 2014.
18. Brungart, D.S.; Kruger, S.E.; Kwiatkowski, T.; Heil, T.; Cohen, J. The effect of walking on auditory localization, visual discrimination, and aurally aided visual search. *Hum, Factors* **2019**, *61*, 976–991. [CrossRef]
19. Ruediger, P.; Spilski, J.; Kartal, N.; Gsuck, S.; Beese, N.O.; Schlittmeier, S.J.; Lachmann, T.; Ebert, A. Cognitive indicators for acoustic source localization and presence in a vivid 3D scene. In Proceedings of the 23rd International Congress on Acoustics, Aachen, Germany, 9–13 September 2019.

20. Hart, S.G.; Staveland, L.E. Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Adv. Psychol.* **1988**, *52*, 139–183. [CrossRef]
21. Schubert, T.; Friedmann, F.; Regenbrecht, H. The experience of presence: Factor analytic insights. *Presence Teleoperators Virtual Environ.* **2001**, *10*, 266–281. [CrossRef]
22. Beatty, J. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychol. Bull.* **1982**, *91*, 276–292. [CrossRef]
23. Beatty, J.; Lucero-Wagoner, B. The pupillary system. In *Handbook of Psychophysiology*; Cacioppo, J.T., Tassinary, L.G., Berntson, G.G., Eds.; Cambridge University Press: Cambridge, UK, 2000; pp. 142–162.
24. Chen, S.; Epps, J. Using Task-Induced Pupil Diameter and Blink Rate to Infer Cognitive Load. *Hum. Comput. Interact.* **2014**, *29*, 390–413. [CrossRef]
25. Mathot, S. Pupillometry: Psychology, Physiology, and Function. *J. Cogn.* **2018**, *1*, 1. [CrossRef] [PubMed]
26. de Greef, T.; Lafeber, H.; van Oostendorp, H.; Lindenberg, J. Eye movement as indicators of mental workload to trigger adaptive automation. In *Foundations of Augmented Cognition. Neuroergonomics and Operational Neuroscience. FAC 2009. Lecture Notes in Computer Science*; Schmorrow, D., Estabrooke, I., Grootjen, M., Eds.; Springer: Berlin/Heidelberg, Germany, 2009; Volume 5638, pp. 219–228.
27. Palinko, O.; Kun, A. Exploring the influence of light and cognitive load on pupil diameter in driving simulator studies. In Proceedings of the Symposium on Eye Tracking Research and Applications (ETRA'12), Santa Barbara, CA, USA, 28–30 March 2012; Association for Computing Machinery: New York, NY, USA, 2012.
28. Benedetto, S.; Pedrotti, M.; Minin, L.; Baccino, T.; Re, A.; Montanari, R. Driver workload and eye blink duration. *Transp. Res. Part F Traf. Psych. Behav.* **2011**, *14*, 199–208. [CrossRef]
29. Zheng, B.; Jiang, X.; Atkins, M.S. Detection of Changes in Surgical Difficulty: Evidence from Pupil Responses. *Surg. Innov.* **2015**, *22*, 629–635. [CrossRef]
30. Veltman, J.A.; Gaillard, A.W.K. Physiological workload reactions to increasing levels of task difficulty. *Ergon* **1998**, *41*, 656–669. [CrossRef] [PubMed]
31. Recarte, M.; Perez, E.; Conchillo, A.; Nunes, L. Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *Span. J. Psychol.* **2008**, *11*, 374–385. [CrossRef]
32. Merat, N.; Jamson, A.H.; Lai, F.C.H.; Carsten, O. Highly Automated Driving, Secondary Task Performance, and Driver State. *Hum. Factors* **2012**, *54*, 762–771. [CrossRef] [PubMed]
33. Said, S.; Gozdzik, M.; Roche, T.R.; Braun, J.; Rössler, J.; Kaserer, A.; Spahn, D.R.; Nöthiger, C.B.; Tscholl, D.W. Validation of the Raw National Aeronautics and Space Administration Task Load Index (NASA-TLX) Questionnaire to Assess Perceived Workload in Patient Monitoring Tasks: Pooled Analysis Study Using Mixed Models. *J. Med. Internet. Res.* **2020**, *22*, e19472. [CrossRef]
34. Hershman, R.; Henik, A.; Cohen, N. A novel blink detection method based on pupillometry noise. *Behav. Res. Methods* **2018**, *50*, 107–114. [CrossRef] [PubMed]
35. Olk, B.; Dinu, A.; Zielinski, D.J.; Kopper, R. Measuring visual search and distraction in immersive virtual reality. *R. Soc. Open Sci.* **2018**, *5*, 172331. [CrossRef]
36. Hekele, F.; Spilski, J.; Bender, S.; Lachmann, T. Remote vocational learning opportunities—A comparative eye-tracking investigation of educational 2D videos versus 360° videos for car mechanics. *Br. J. Educ. Technol.* **2021**. [CrossRef]
37. Marquart, G.; Cabrall, C.; de Winter, J. Review of eye-related measures of drivers' mental workload. *Procedia Manuf.* **2015**, *3*, 2854–2861. [CrossRef]
38. Zheng, B.; Jiang, X.; Tien, G.; Meneghetti, A.; Panton, O.N.; Atkins, M.S. Workload assessment of surgeons: Correlation between NASA TLX and blinks. *Surg. Endosc.* **2012**, *26*, 2746–2750. [CrossRef]
39. Berguer, R.; Smith, W.; Chung, Y. Performing laparoscopic surgery is significantly more stressful for the surgeon than open surgery. *Surg. Endosc.* **2001**, *15*, 1204–1207. [CrossRef]
40. Ahlstrom, U.; Friedman-Berg, F. Using eye movement activity as a correlate of cognitive workload. *Intern. J. Indust. Ergon.* **2006**, *36*, 623–636. [CrossRef]
41. Kim, J.; Sunil Kumar, Y.; Yoo, J.; Kwon, S. Change of Blink Rate in Viewing Virtual Reality with HMD. *Symmetry* **2018**, *10*, 400. [CrossRef]
42. Mathur, A.; Gehrmann, J.; Atchison, D.A. Pupil shape as viewed along the horizontal visual field. *J. Vis.* **2013**, *13*, 6. [CrossRef] [PubMed]
43. Duchowski, A.T.; Krejtz, K.; Krejtz, I.; Biele, C.; Niedzielska, A.; Kiefer, P.; Raubal, M.; Giannopoulos, I. The Index of Pupillary Activity: Measuring Cognitive Load *vis-à-vis* Task Difficulty with Pupil Oscillation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems—CHI '18, Montréal, QC, Canada, 21–26 April 2018.